Semantics-Preserved Data Augmentation for Aspect-Based Sentiment Analysis

Ting-Wei Hsu [1], **Chung-Chi Chen** [1], **Hen-Hsen Huang** [2,3], **Hsin-Hsi Chen** [1,3]
[1] Department of Computer Science and Information Engineering, National Taiwan University, Taiwan
[2] Institute of Information Science, Academia Sinica, Taiwan
[3] MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan
{twhsu,cjchen}@nlg.csie.ntu.edu.tw, hhhuang@iis.sinica.edu.tw, hhchen@ntu.edu.tw

Abstract

Both the issues of data deficiencies and semantic consistency are important for data augmentation. Most of previous methods address the first issue, but ignore the second one. In the cases of aspect-based sentiment analysis, violation of the above issues may change the aspect and sentiment polarity. In this paper, we propose a semantics-preservation data augmentation approach by considering the importance of each word in a textual sequence according to the related aspects and sentiments. We then substitute the unimportant tokens with two replacement strategies without altering the aspect-level polarity. Our approach is evaluated on several publicly available sentiment analysis datasets and the real-world stock price/risk movement prediction scenarios. Experimental results show that our methodology achieves better performances in all datasets.

1 Introduction

Data annotation, the first step of most artificial intelligence researches, often takes lots of time and is very expensive. Many studies propose methodologies for augmenting data based on a few annotations (Wei and Zou, 2019; Jiao et al., 2020; Xie et al., 2020; Wu et al., 2019; Kobayashi, 2018) to reduce the cost of annotation. However, most of them encounter a problem—hardly ensuring readability and semantic coherence. To overcome these problems, we introduce a novel method, selective perturbed masking (SPM), to measure the importance of each word in a textual sentence, and further replace the unimportant word with pre-trained language models. In this way, the semantics of a given instance will be preserved, and the training set will be enlarged with various auto-generated data.

While testing the data augmentation methods, classification tasks are often adopted. Sentiment analysis is one of the well-known classification tasks. To mining more fine-grained information, aspect-based sentiment analysis (ABSA) is proposed. ABSA not only aims at detecting the sentiment polarity but also attempts to mine the analysis aspect. It can be extended to subtasks of aspect category sentiment classification (ACSC), aspect term sentiment classification (ATSC), and aspect term extraction (ATE). For example, the answers to these tasks to the given sentence "the staff was so kind" are (service, positive), (staff, positive), and "staff", respectively. These tasks are commonly taken as examples for evaluating the performance of augmentation methods. In this paper, we explore both sentiment analysis and all subtasks in ABSA to show the usefulness of the proposed method in semantics preservation.

In addition to probe on sentiment analysis and ABSA tasks, we also experiment on stock price and risk movement prediction tasks. These experiments have always been considered as real-world sentiment analysis application scenarios (Xu and Cohen, 2018). The experimental results on all kinds of datasets support the usefulness of the proposed data augmentation method, and also indicate that the improvement of using the proposed method is larger than that of using other augmentation methods. Our main contributions are summarized as follows:

1. We propose a semantics-preserved augmentation method,[1] which provides more training data without changing the meaning of the given instances in augmentation.

2. Our experimental results show the proposed method can achieve better performances in four aspect-based sentiment analysis datasets and two sentiment analysis datasets.

3. An additional exploration of a real-world scenario on stock price/risk movement predic-

[1] https://github.com/Quant-NLP/SPDAug-ABSA

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 4417–4422
November 7–11, 2021.

tion supports the robustness of the proposed method.

2 Related Work

2.1 Data Augmentation

Jiao et al. (2020) utilize the word embeddings to obtain the similar terms as the substitutions. Wei and Zou (2019) add noises by randomly adding, deleting, swapping the words, and substituting words with synonyms. Kobayashi (2018); Wu et al. (2019) consider label information and use the language model to randomly replacing single-word with more diverse substitutions. Xie et al. (2020) replace the uninformative words based on TF-IDF scores. Although previous works show that their methods can improve the performance of some NLP tasks, their data augmentation methods may change the meanings of the given instances. In this paper, we propose a method that not only provides more data, but also retains the meanings of the original instances. Our experimental results show that the proposed method performs well in all datasets.

2.2 Aspect-Based Sentiment Analysis

Wang et al. (2016) propose an attention-based LSTM, which can concentrate on distinct parts of a sentence by calculating the corresponding attention weights, to learn aspect embedding. Ma et al. (2017) find that interactive attention networks can learn the representations of target and context separately, which is helpful to sentiment classification. BERT-based (Devlin et al., 2019) methods have shown to be effective on ABSA (Hoang et al., 2019; Xu et al., 2019). Because BERT-based methods achieve the best performance in most ABSA tasks, we adopt BERT as the base architecture for test performance of data augmentation methods.

3 Methods

This section describes how we combine an auxiliary sentence with SPM in detail to decide which words are unimportant to the aspect or sentiment of a given instance. After deciding the unimportant words, we use the proposed token replacement methods to construct the augmented sentences, in which aspect and sentiment are not altered.

3.1 An Auxiliary Sentence Approach

Inspired by Sun et al. (2019) and Schick and Schütze (2021), we utilize an auxiliary sentence containing the aspect and the sentiment in a review. For each review, we concatenate on the auxiliary sentence and the review with a special token [**SEP**]. For example, the auxiliary sentence "the polarity of the service is positive" and the review "the staff was so kind to us" are concatenated to "the polarity of the service is positive [**SEP**] the staff was so kind to us". An input sentence (S) is thus formulated as follows: $S = [w^a_1, ..., w^a_{aspect}, ..., w^a_{sentiment}, w_{[\mathbf{SEP}]}, w^r_1 ..., w^r_N]$, where w^a and w^r denote words in auxiliary sentence and review, respectively. w_{aspect} and $w_{sentiment}$ are the words denoting aspect and sentiment, respectively, and N is the length of the review.

3.2 Selective Perturbed Masking (SPM)

Wu et al. (2020) introduce perturbed masking (PM) to analyze syntactic information, and verify its effectiveness on syntactic parsing and discourse dependency parsing. In this work, we propose selective perturbed masking (SPM) to estimate the correlation between the tokens in reviews and sentiment words (aspect terms) in the auxiliary sentence. The following procedure is proposed to measure the impact $w^r_i (1 \leq i \leq N)$ on predicting w^a_{aspect} and $w^a_{sentiment}$, respectively. First, we replace w_{aspect} ($w_{sentiment}$) with the special token [**MASK**], and use this word sequence as BERT's input for predicting the masked word. The output embedding is called E^a. Note that the SPM method that masks w_{aspect} is named AS-SPM, and the SPM method that masks $w_{sentiment}$ is called Senti-SPM. Then, we replace both w_{aspect} ($w_{sentiment}$) and $w^r_i (1 \leq i \leq N)$ with special token [**MASK**]. The BERT's output embedding at the position of w_{aspect} ($w_{sentiment}$) is considered as E^r_i. Thirdly, we calculate the Euclidean distance (ED) between E^a and each E^r_i, where

$$ED(x, y) = ||x - y||_2 \quad (1)$$

Finally, we consider the w^r_j is an unimportant term to w^a_{aspect} ($w^a_{sentiment}$) if ED(E^a,E^r_j) is lower than the averaged ED of all (E^a,E^r_i) pairs ($1 \leq i \leq N$), which is formulated as follows.

$$ED(E^a, E^r_j) < \frac{1}{N} \sum_{i=\mathbf{1}}^{\mathbf{N}} ED(E^a, E^r_i) \quad (2)$$

3.3 Token Replacement Strategy

After doing SPM for reviews, our model replaces the unimportant terms under the following two strategies.

Task	ACSC	ATSC				ATE				SC	
Dataset	Rest14	Rest14	Lap14	Rest15	Rest16	Rest14	Lap14	Rest15	Rest16	MR	SST-2
Train	3,712	3,602	2,313	1,610	2,417	3,041	3,045	1,315	2,000	8,529	6,920
Test	1,025	1,120	638	802	825	800	800	685	676	1,067	1,821
Total	4,737	4,722	2,951	2,412	3,242	3,841	3,845	2,000	2,676	9,596	8,741

Table 1: Statistics of datasets.

Model	ACSC	ATSC				ATE			
	Rest14	Rest14	Lap14	Rest15	Rest16	Rest14	Lap14	Rest15	Rest16
Bert_{base}	$82.98_{0.78}$	$79.48_{0.64}$	$75.32_{1.08}$	$81.62_{1.07}$	$86.58_{0.56}$	$86.44_{0.49}$	$78.49_{1.38}$	$66.10_{4.61}$	$72.42_{2.38}$
+ BT	$82.45_{0.62}$	$79.98_{0.51}$	$75.76_{1.19}$	$82.61_{0.60}$	$86.22_{0.58}$	$86.57_{0.48}$	$80.66_{2.27}$	$70.34_{1.65}$	$74.23_{0.64}$
+ EDA	$82.82_{0.15}$	$79.82_{0.58}$	$76.11_{0.58}$	$81.77_{1.43}$	$85.65_{0.53}$	-	-	-	-
+ C-BERT	$83.45_{1.14}$	$79.67_{0.80}$	$76.45_{0.90}$	$80.37_{2.56}$	$85.57_{1.69}$	$86.73_{0.15}$	$81.00_{1.68}$	$69.21_{1.14}$	$75.19_{0.57}$
+ AS-SPM & AE	$83.14_{0.98}$	$80.55_{0.42}$	$76.33_{1.19}$	$83.91_{0.98}$	$\mathbf{87.85}_{0.38}$	$\mathbf{87.18}_{0.63}$	$\mathbf{82.86}_{1.50}$	$\mathbf{70.68}_{1.15}$	$\mathbf{75.62}_{0.64}$
+ Senti-SPM & AE	$84.07_{0.36}$	$80.50_{0.80}$	$77.21_{0.61}$	$84.28_{0.64}$	$87.61_{0.40}$	-	-	-	-
+ AS-SPM & Seq2Seq	$\mathbf{84.17}_{0.94}$	$81.19_{0.65}$	$\mathbf{77.93}_{0.43}$	$\mathbf{84.46}_{0.22}$	$87.55_{0.45}$	$87.04_{0.54}$	$81.51_{1.07}$	$69.27_{0.87}$	$75.24_{0.58}$
+ Senti-SPM & Seq2Seq	$83.39_{1.03}$	$\mathbf{81.50}_{0.47}$	$77.55_{1.31}$	$83.74_{1.25}$	$87.81_{0.54}$	-	-	-	-

Table 2: Experimental results.

Auto Encoding (AE): The AE model is initialized with the pre-trained weights of BERT (Devlin et al., 2019), which will predict the masked word. We use the predicted word to replace the unimportant term.

Sequence-to-Sequence (Seq2Seq): The Seq2Seq model is initialized with the pre-trained weights of BART (Lewis et al., 2020) which includes encoder and decoder. The predicted word from the decoder is adopted for replacing the unimportant term.

4 Experiments

4.1 Experimental Setup

We experiment on four widely-used datasets in ABSA, including Lap14 (Pontiki et al., 2014), Rest14 (Pontiki et al., 2014), Rest15 (Pontiki et al., 2015), and Rest16 (Pontiki et al., 2016). Furthermore, we evaluate our model on two sentiment classification (SC) benchmark datasets including Stanford Sentiment Treebank (SST-2) (Socher et al., 2013) and Movie Review (MR) (Pang and Lee, 2005). The statistics of these datasets are reported in Table 1.

In our experiments, we use the *BERT-base-uncased* model to show the performances with and without the proposed augmentation methods. Additionally, we compare with commonly used data augmentation methods, including Back Translation (BT) (Edunov et al., 2018), Easy Data Augmentation (EDA) (Wei and Zou, 2019), and C-BERT (Wu et al., 2019). In ACSC, ATSC, and SC tasks, we double the original training set in size. In ATE task, we augment the reviews according to the number of aspect terms. Accuracy is adopted as the evaluation metric for ACSC, ATSC, and SC tasks. F1-score is used in ATE task.

4.2 Experimental Results

We report the averaged results across five random seeds in Table 2, and the standard deviations are also shown in subscripts. We do not adopt EDA to the ATE task because the insertion and the deletion operations are not suitable for token-level tasks. Firstly, we find that the combinations of the SPM settings (AS-SPM and Senti-SPM) and token replacement strategies (AE and Seq2Seq) achieve better performances on all settings with stable results (lower standard deviations). That indicates our augmentation methods are effective. Secondly, some approaches slightly harm the performance of some datasets. For example, using BT and EDA in ACSC-Rest14 and ATSC-Rest16 gets lower performances than using vanilla BERT; using C-BERT in ATSC-Rest15 and ATSC-Rest16 gets lower performances than using vanilla BERT. Additionally, the proposed SPM consistently outperforms the random masking strategy (C-BERT). Thirdly, the proposed token replacement strategy Seq2Seq performs well in ACSC and ATSC, and AE achieves the best results in ATE.

5 Discussion

5.1 Multilingual Experiment

In this section, we utilize Google Translate to translate corresponding auxiliary sentences, and experiment on ABSA datasets (Pontiki et al., 2016) in

Model	Language						
	AR	CH	DU	FR	RU	ES	TU
Bert$_{base}$	$88.48_{0.78}$	$93.79_{0.87}$	$85.37_{3.36}$	$85.98_{2.47}$	$90.78_{1.86}$	$81.66_{1.17}$	$66.81_{1.73}$
+ BT	$88.24_{0.87}$	$94.58_{1.40}$	$\mathbf{88.20}_{1.66}$	$87.66_{3.93}$	$93.90_{1.50}$	$84.00_{1.70}$	$72.89_{3.92}$
+ C-BERT	$87.88_{2.24}$	$94.20_{1.58}$	$84.74_{2.82}$	$88.41_{1.69}$	$92.96_{1.46}$	$80.16_{2.72}$	$71.59_{6.34}$
+ AS-SPM & AE	$89.20_{1.01}$	$95.48_{1.17}$	$86.32_{1.30}$	$88.41_{1.56}$	$\mathbf{94.21}_{0.42}$	$\mathbf{84.66}_{1.51}$	$72.31_{2.13}$
+ Senti-SPM & AE	$87.41_{0.88}$	$94.31_{0.32}$	$86.16_{1.54}$	$88.97_{1.02}$	$92.81_{1.28}$	$83.00_{2.32}$	$71.73_{4.06}$
+ AS-SPM & Seq2Seq	$\mathbf{90.28}_{1.34}$	$95.31_{0.74}$	$86.63_{1.25}$	$\mathbf{89.15}_{1.41}$	$93.75_{1.23}$	$82.83_{1.51}$	$71.44_{2.74}$
+ Senti-SPM & Seq2Seq	$88.48_{1.76}$	$\mathbf{95.58}_{0.71}$	$87.26_{1.95}$	$88.78_{2.19}$	$91.40_{0.95}$	$83.66_{2.80}$	$\mathbf{73.62}_{3.01}$

Table 3: Experimental results on other languages.

Model	MR	SST-2
Bert$_{base}$	$85.64_{0.77}$	$90.39_{0.81}$
+ BT	$85.90_{0.37}$	$90.82_{0.54}$
+ EDA	$85.54_{0.41}$	$90.53_{0.88}$
+ C-BERT	$85.02_{1.38}$	$90.16_{0.46}$
+ Senti-SPM & AE	$85.75_{0.50}$	$90.91_{0.34}$
+ Senti-SPM & Seq2Seq	$\mathbf{86.52}_{0.59}$	$\mathbf{91.51}_{0.42}$

Table 4: Experimental results on sentiment classification.

Model	MAMS	
	ATSC	ACSC
Bert$_{base}$	$82.23_{0.41}$	$73.45_{1.38}$
+ BT	$82.73_{0.37}$	$73.60_{1.02}$
+ EDA	$82.78_{0.20}$	$74.53_{1.40}$
+ C-BERT	$82.34_{0.48}$	$74.22_{0.92}$
+ AS-SPM & AE	$82.09_{0.41}$	$75.29_{0.93}$
+ Senti-SPM & Seq2Seq	$82.33_{0.72}$	$73.89_{1.08}$
+ AS-SPM & AE	$83.00_{0.80}$	$\mathbf{76.15}_{0.79}$
+ Senti-SPM & Seq2Seq	$\mathbf{83.17}_{0.51}$	$75.27_{0.69}$

Table 5: Experimental results on MAMS datasets.

other languages, including Arabic (AR), Chinese (CH), Dutch (DU), French (FR), Russian (RU), Spanish (ES), and Turkish (TU). The experimental results are shown in Table 3. In most languages, the proposed method improves more performance than other data augmentation methods.

5.2 Multi-Aspect Multi-Sentiment Experiment

Jiang et al. (2019) propose a multi-aspect multi-sentiment dataset, MAMS. In MAMS, each instance is annotated with different sentiments from at least two aspects. They claim that this is a more challenging dataset than that in the previous works. We further experiment on this dataset with the proposed method, and report the results in Table 5. These results support that the proposed method is also helpful in this more challenging dataset in both ATSC and ACSC tasks.

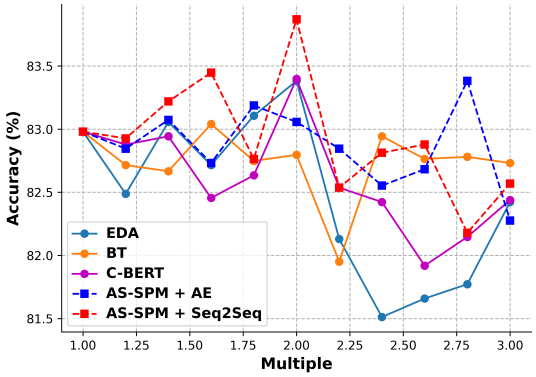

Figure 1: Results under different training set size.

5.3 Influence of Augmentation Size

Figure 1 shows the results on ACSC-Rest14 under different multiples of the training set size. The performances of models become better than that only using the original training set when the multiple is between 1 and 2. It also shows that using too much training data generated by the proposed method harms the performance of ABSA because it may contain too much noise.

5.4 Case Study

Table 6 presents the augmented results of different approaches. It shows that previous approaches may change the aspect or sentiment of an instance. For example, EDA generates an unmeaningful sentence. Although "ugly" could be a hint for negative sentiment, the aspect of the generated instance is changed. C-BERT shows the other worse case—the sentiment of the generated review is changed. Both cases show the importance of the proposed idea, i.e., controlling the aspect or sentiment word when generating a new sentence. In contrast, the proposed approach is controllable. When using

Original review:	But the staff was so horrible to us.
BT	But the staff were so awful for us.
EDA	But so staff was the ugly to uranium.
C-BERT	But the situation was being good to me.
AS-SPM & AE	But the staff was always horrible to me.
Senti-SPM & AE	But the situation was so horrible to me.

Table 6: Examples of different approaches. The colored words are changed by the designated approaches.

AS-SPM, the aspect will not be changed. On the other hand, when using Senti-SPM, the proposed approach keeps the same sentiment polarity as the original review.

5.5 Stock Price/Risk Movement Prediction

In financial markets, return and risk are two aspects that most investors focus on. The task setting is similar to ABSA tasks. In this section, we discuss the experiments of stock price/risk movement prediction in the benchmark dataset, StockNet (Xu and Cohen, 2018). We ask models to predict whether the return (risk) will increase or decrease after n days, where return (μ_t) is $\frac{P_t - P_{t-n}}{P_{t-n}}$ and P_t is the close price at time t. Risk (σ_t) is defined as follows:

$$\sigma_t = \sqrt{\frac{1}{n-1}\sum_{t=1}^{n}(\mu_t - \frac{1}{n}\sum_{t=1}^{n}\mu_t)^2} \quad (3)$$

where n is 3. Total 19,107 instances are included in the StockNet, and we use 85% of instances as training set and the rest as test set. Accuracy (ACC.) is adopted as the evaluation metric.

The auxiliary sentences in this experiment are different from those in ABSA tasks. For example, when predicting risk movement, the auxiliary sentence is " Market risk will [**MASK**] ". In accordance with the task, there are two kinds of SPM, i.e., Return-SPM and Risk-SPM. Table 7 shows the experimental results. The proposed methods outperform other data augmentation methods in both price/risk movement prediction tasks.

5.6 Influence of Auxiliary Sentence

In this section, we discuss an interesting research question: whether different auxiliary sentences will influence the performance. For example, how good the performance is if we only use simple "[MASK]" as the auxiliary sentence, and how tense influences performance. We use the experiments on stock risk movement prediction as an example. Table 8 presents our pilot results for the research question. The experimental results show that adding aspect term ("Risk") performs better than using simple "[MASK]" tag only. Additionally, the tense of the auxiliary sentence is also influential. Since "risk" may be related to different issues, we find that adding an issue-specific term ("Market") can provide slight improvement. In sum, our experiments show the importance of selecting auxiliary sentences in the data augmentation process.

Model	Aspect	
	Return	Risk
Bert$_{base}$	$50.57_{2.99}$	$50.74_{3.36}$
+ BT	$51.24_{0.95}$	$51.55_{3.42}$
+ EDA	$51.84_{1.83}$	$51.45_{1.77}$
+ C-BERT	$52.13_{1.25}$	$51.68_{2.66}$
+ Return-SPM & AE	$52.87_{0.78}$	$53.91_{1.46}$
+ Return-SPM & Seq2Seq	$\mathbf{54.04}_{0.77}$	$53.17_{2.60}$
+ Risk-SPM & AE	$51.98_{0.68}$	$55.07_{2.47}$
+ Risk-SPM & Seq2Seq	$52.02_{0.54}$	$\mathbf{55.32}_{2.96}$

Table 7: Experimental results on StockNet.

Auxiliary Sentence	ACC.
[MASK]	$50.57_{2.68}$
Risk [MASK]	$51.51_{2.56}$
Risk will [MASK]	$52.12_{2.82}$
Market risk will [MASK]	$\mathbf{55.32}_{2.96}$

Table 8: Results using different auxiliary sentences.

6 Conclusion

In this paper, we present a controllable augmentation for ABSA, which is controllable to generate reasonable reviews without converting aspect-level polarity. We propose SPM to measure the impact of the related words on deciding specific aspect and sentiment, and adopt two replacement strategies to ABSA tasks. Experimental results show the effectiveness and robustness of our approaches. Additionally, the exploration in the financial application scenario also supports the usefulness of the proposed method. In the future, we plan to use the proposed method for data augmentation on longer documents and for generating the training instances of low-resource languages.

Acknowledgments

This research was partially supported by Ministry of Science and Technology, Taiwan, under grants MOST 109-2218-E-009-014, MOST 110-2634-F-002-028, and MOST 110-2221-E-002 -128 -MY3.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *EMNLP*.

Mickel Hoang, Oskar Alija Bihorac, and Jacobo Rouces. 2019. Aspect-based sentiment analysis using bert. In *NoDaLiDa*.

Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. 2019. A challenge dataset and effective models for aspect-based sentiment analysis. In *EMNLP*, pages 6280–6285.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. Tinybert: Distilling bert for natural language understanding. In *EMNLP*.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *NAACL*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. In *IJCAI*.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *ACL*.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *SemEval-2016*.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 task 12: Aspect based sentiment analysis. In *SemEval 2015*.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *SemEval*.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. *NAACL*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In *NAACL*.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based lstm for aspect-level sentiment classification. In *EMNLP*.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *EMNLP-IJCNLP*.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional bert contextual augmentation. In *ICCS*.

Zhiyong Wu, Yun Chen, Ben Kao, and Qun Liu. 2020. Perturbed masking: Parameter-free probing for analyzing and interpreting bert. In *ACL*.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. In *NeurIPS*.

Hu Xu, Bing Liu, Lei Shu, and S Yu Philip. 2019. Bert post-training for review reading comprehension and aspect-based sentiment analysis. In *NAACL*.

Yumo Xu and Shay B. Cohen. 2018. Stock movement prediction from tweets and historical prices. In *ACL*.